Entrez Help

Created: January 20, 2006
Updated: April 19, 2006

Entrez integrates the scientific literature, DNA and protein sequence databases, 3D protein structure and protein domain data, population study datasets, expression data, assemblies of complete genomes, and taxonomic information into a tightly interlinked system. It is a retrieval system designed for searching its linked databases. Help using the literature component of Entrez, known as PubMed, is also available. Go to PubMed Help.

Entrez, the Life Sciences Search Engine : Global Query

The Entrez page is home to the Entrez Global Query database search engine (the Entrez cross-database search page). The entire group of individual Entrez databases is organized on this page with literature databases at the top including PubMed, PubMed Central, Journals, Books, OMIM and OMIA. The NCBI Site Search is also listed. The sequence databases include Nucleotide, Protein, Genome, Structure, and SNPs. The remaining databases are Taxonomy, Gene, UniGene, HomoloGene, Conserved Domains, 3D Domains, UniSTS, PopSet, GEO Profiles, GEO Datasets, PubChem Bio-Assay, PubChem Compound, PubChem Substance, Cancer Chromosomes, Probe, MeSH, Journals and NLM Catalog. Links to popular NCBI Web pages, such as PubMed, Human Genome, Map Viewer, and BLAST, are on the toolbar. There is also a link to the "GenBank" database, leading to the Nucleotide database.

By using the Entrez Global query, a search across all Entrez databases is performed by entering a simple search term or phrase in the "Search across databases" query box. Select the Go button to execute the search, or press the Enter button on your keyboard. The CLEAR button erases search terms in the query box; use it to begin a new search. The results found in each database are displayed on the Global Query page. Click on the result number or its adjacent database name to get to the specific results. See the link to the Global Query Help document, which is to the right of the CLEAR button.

The Databases

Umbrella Nucleotide Database

NCBI's traditional Nucleotide database is now also searchable as three component databases: "EST" (containing EST sequences), "GSS" (containing GSS sequences) and "CoreNucleotide" (which comprises the remaining nucleotide sequences). This will allow faster searching and more specific results for nucleotide sequence records. When a search is done in the Nucbleotide database, Entrez search results are also shown for the three component Nucleotide databases on the Search statistic line. The component Nucleotide databases together contain all

the sequence data from GenBank, EMBL, and DDBJ, the members of the International Collaboration of Sequence Databases. Link to Entrez Help document for more details on the databases and how to search them.

As an example of the change in the format of results returned from an Entrez Nucleotide search, consider a simple search using a query of "mouse[organism]". The result are shown with links to the split of data in the GenBank EST division, the GenBank GSS division, and the CoreNucleotide subset. Users can select results from the specific nucleotide dataset of interest by clicking on the appropriate results link.

The new component databases are included within the Entrez linking scheme and Links within and between databases can be selected as usual from the various datasets. Popular search strategies such as the Limits, Preview/Index, History, and MyNCBI can be used within each individual database.

Specialized search fields are available for the each new Nucleotide component database and can be seen in the respective indexed search fields in the "Add Term(s) to Query or View Index:" section of the Preview/Index tab for each. New fields available for the EST component database include EST ID, EST Name, and Library. Within the GSS component database, new fields include GSS ID, GSS Name, Library Class, and Library Name. CoreNucleotide contains the same 23 search fields as the traditional Nucleotide database.

All previous Entrez functionality remains, such as Clipboard where searches can be be saved temporarily and MyNCBI where searches can be indefinitely saved to run at user-selected intervals. Details and History pages are available for the individual search sets. To see search Details, the database of interest must be selected when performing a search from the main Nucleotide page.

Patent sequences are incorporated through arrangements with the U.S. Patent and Trademark Office (USPTO) and via the collaborating international databases from other international patent offices.

The Umbrella Nucleotide database also includes the Reference Sequence (RefSeq) records. RefSeqs are an NCBI-curated non-redundant set of sequences. Click on this link for more information about the RefSeq [http://www.ncbi.nlm.nih.gov/RefSeq] project.

Protein Database

The Protein database contains sequence data from the translated coding regions from DNA sequences in GenBank, EMBL, and DDBJ as well as protein sequences submitted to Protein Information Resource (PIR), SWISS-PROT, Protein Research Foundation (PRF), and Protein Data Bank (PDB) (sequences from solved structures).

Genome Database

The Genome database provides views for a variety of genomes, complete chromosomes, sequence maps with contigs, and integrated genetic and physical maps.

Structure Database

The Structure database or Molecular Modeling Database (MMDB) contains experimental data from crystallographic and NMR structure determinations. The data for MMDB are obtained from the Protein Data Bank (PDB). The NCBI has cross-linked structural data to bibliographic information, to the sequence databases, and to the NCBI taxonomy.

Use Cn3D [http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml], the NCBI 3D structure viewer, for easy interactive visualization of molecular structures from Entrez.

3D Domains

3D Domains contains protein domains from the Entrez Structure Database. See 3D Domains.

Conserved Domains

Conserved Domains is a database of protein domains. The source databases for Conserved Domains are Pfam, Smart, and COG. See CDD Help [http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd_help.shtml].

UniSTS

UniSTS is a unified, non-redundant view of sequence tagged sites (STSs). UniSTS integrates marker and mapping data from a variety of public resources. Data sources include dbSTS, RHdb, GDB, various human maps (Genethon genetic map, Marshfield genetic map, Whitehead RH map, Whitehead YAC map, Stanford RH map, NHGRI chr 7 physical map, and WashU chrX physical map), and various mouse maps (Whitehead RH map, Whitehead YAC map, and Jackson Laboratory's MGD map). See UniSTS.

Gene

Gene provides a unified query environment for genes defined by sequence and/or in NCBI's Map Viewer. You can query on names, symbols, accessions, publications, GO terms, chromosome numbers, E.C. numbers, and many other attributes associated with genes and the products they encode. See Gene and Gene Help.

UniGene

UniGene is an experimental system for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters. Each UniGene cluster contains sequences that represent a unique gene, as well as related information such as the tissue types in which the gene has been expressed and map location. See UniGene and its Query Tips and FAQs.

HomoloGene

HomoloGene is a system for automated detection of homologs among eukaryotic gene sets. See [http://www.ncbi.nlm.nih.gov/projects/gensat/]HomoloGene and its Query Tips.

SNP

SNP is a central repository database for both single-base nucleotide substitutions and short deletion and insertion polymorphisms. For the search page and available search fields and search examples, see SNP [http://www.ncbi.nlm.nih.gov/SNP].

PopSet Database

The PopSet database contains aligned sequences submitted as a set resulting from a population, phylogenetic, or mutation study. These alignments describe such events as evolution and population variation. The PopSet database contains both nucleotide and protein sequence data. See PopSet.

Taxonomy Database

The Taxonomy database contains the names of all organisms that are represented in the NCBI genetic database by at least one nucleotide or protein sequence. For the context of the Taxonomy database see Taxonomy and Taxonomy FAQs.

GEO Profiles

This database stores individual gene expression and molecular abundance profiles assembled from the Gene Expression Omnibus (GEO) repository. See the GEO QuickQuery and FAQs [http://www.ncbi.nlm.nih.gov/geo/info/qqtutorial.html].

GEO Datasets

This database stores curated gene expression and molecular abundance DataSets assembled from the Gene Expression Omnibus (GEO) repository. See the GEO FAQs.

GENSAT

The GENSAT (Gene Expression Nervous System Atlas) project aims to map the expression of genes in the central nervous system of the mouse, using both in situ hybridization and transgenic mouse techniques. See the the GENSAT [http://www.ncbi.nlm.nih.gov/projects/gensat] homepage to search and GENSAT FAQ [http://www.ncbi.nlm.nih.gov/projects/gensat/static/faq.shtml] for additional information.

Cancer Chromosomes

Cancer Chromosomes contains three cancer cytogenetic databases: the NCI Mitelman Database of Chromosome Aberrations in Cancer, the NCI Recurrent Chromosome Aberrations in Cancer, and the NCI and NCBI SKY/M-FISH & CGH Database. Karyotype, SKY/M-FISH, and CGH data can be searched simultaneously. Similarity searches demonstrate cytogenetic and clinical relatedness at varying levels of specificity. See the Cancer Chromosomes Web site to search and for additional information

PubChem Compound

The PubChem Compound Database contains validated chemical depiction information provided to describe substances in PubChem Substance. See the PubChem Compound Web site to search and for additional information.

PubChem Substance

The PubChem Substance Database contains descriptions of chemical samples, from a variety of sources, and links to PubMed citations, protein 3D structures, and biological screening results that are available in PubChem BioAssay. See the PubChem Substance Web site to search or for additional information.

PubChem BioAssay

The PubChem BioAssay Database contains bioactivity screens of chemical substances described in PubChem Substance. It provides searchable descriptions of each bioassay, including descriptions of the conditions and readouts specific to that screening procedure. See the PubChem BioAssay Web site to search and for additional information.

PubMed Central

PubMed Central (PMC) is the U.S. National Library of Medicine's digital archive of life sciences journal literature. Access to the full text of articles in PMC is free, except where a journal requires a subscription for access to recent articles. See PubMed Central, the PubMed Central Help, and PubMed Central FAQs [http://www.pubmedcentral.gov/about/faq.html].

Journals

The Journals database can be searched using the journal title, MEDLINE abbreviation, NLM ID, ISO abbreviation, or ISSN. The database includes the journals in all Entrez databases, e.g., PubMed, Nucleotide, Protein. See Journals.

MeSH

MeSH (Medical Subject Headings) is the National Library of Medicine's controlled vocabulary used for indexing articles in PubMed. MeSH terminology provides a consistent way to retrieve information that may use different terminology for the same concepts. See MeSH.

Bookshelf

The Bookshelf has a collection of Biomedical books that are linked in Entrez. The NCBI Handbook is also available from the Bookshelf. See the Bookshelf and the Books FAQs.

OMIM Database

The OMIM (Online Mendelian Inheritance in Man) database is a catalog of human genes and genetic disorders. See OMIM and OMIM Help [http://www.ncbi.nlm.nih.gov/Omim/omimhelp.html].

OMIA Database

Online Mendelian Inheritance in Animals (OMIA) is a database of genes, inherited disorders and traits in animal species (other than human and mouse) authored by Professor Frank Nicholas of the University of Sydney, Australia, with help from many people over the years. The database contains textual information and references, as well as links to relevant records from OMIM, PubMed, Gene, and soon to NCBI's Phenotype database. See the OMIA Web site.

Probe Database

Probe is a public registry of nucleic acid reagents designed for use in a wide variety of biomedical research applications, together with information on reagent distributors, probe effectiveness, and computed sequence similarities. See Probe and Probe Query Tips

.

Searching Entrez using Global Query

On the Global Query page Entrez cross-database search page, type one or more search terms into the search box. The results for each database are shown and may be selected for the desired database. The Global query page is shown below.

Global Query page

Global Query searches can also be conducted from the NCBI the NCBI homepage [http://www.ncbi.nlm.nih.gov] by entering search terms in the search field. A single database may be selected from the Search All Databases selection menu located on the NCBI homepage [http://www.ncbi.nlm.nih.gov].

More complex search strategies are performed using Boolean operators, and combinations of one of more search field limits.

Database Neighbors and Interlinking

What makes Entrez more powerful than many services is that most of its records are linked to other records, both within a given database (such as Nucleotide) and between databases. Links within a database are called “neighbors” (e.g., Nucleotide neighbors).

Links between databases are also possible. Protein and Nucleotide neighbors are determined by performing similarity searches using the BLAST algorithm to compare the entry amino acid or DNA sequence to all other amino acid or DNA sequences in the database. Nucleotide sequence

records in the Nucleotide database are linked to the PubMed citation of the article in which the sequences were published. Protein sequence records are linked to the nucleotide sequence from which the protein was translated.

See Displaying and Saving Results for more information on links within and between databases.

Boolean Operators

Boolean Operators used in Entrez are:

AND: To 'AND' two search terms together instructs Entrez to find all documents that contain BOTH terms

OR: To 'OR' two search terms together instructs Entrez to find all documents that contain EITHER term.

NOT: To 'NOT' two search terms together instructs Entrez to find all documents that contain search term 1 BUT NOT search term 2.

The Entrez search rules and syntax for using Boolean operators are:

1. Boolean operators AND, OR, NOT must be entered in UPPERCASE (e.g., promoters OR response elements).

2. Entrez processes all Boolean operators in a left-to-right sequence. The order in which Entrez processes a search statement can be changed by enclosing individual concepts in parentheses. The terms inside the parentheses are processed first as a unit and then incorporated into the overall strategy. For example, the search statement: g1p3 AND (response element OR promoter) is processed by Entrez by ORing the terms response element OR promoter first and then ANDing the resulting set of documents with g1p3.

3. Click on the Details button to see how Entrez translated and executed your search strategy.

4. See Writing Advanced Search Statements for more information on using Boolean Operators and Entrez Search Field Qualifiers.

The use of parentheses can change your search results significantly. Compare the number of records retrieved in the Nucleotide database as of February 2006 in each case below.

Example
g1p3 AND (response element OR promoter) retrieves three records
g1p3 AND response element OR promoter retrieves 354,554 records.

Adjacency Searching or Phrase Searching

Subject terms are automatically combined (ANDed). The following query, which is not enclosed in double quotation marks: 16S RNA retrieves all records with the terms 16S AND RNA. See Boolean Operators for more information on combining terms with Boolean Operators.

To closely approximate adjacency term searching or force Entrez to search for a phrase, enter double quotes (" ") around the phrase. The terms in the quoted phrase are searched the order they are placed in the quoted phrase *and* next to each other. This is a “force-phrase” search.

Using quotes forces Entrez to check a phrase list, against which the search terms are matched. It is not true adjacency searching. If the search phrase is not in the phrase list, Entrez returns an informational message "See Details.No items found." and once a nucleotide component database is selected, the message " Quoted phrase not found. See Details.No items found. .

Although phrase searching is useful, it should be used with caution because enclosing search terms in quotes restricts the documents retrieved to only those documents with exact matches to the text string within the quotes. Also, note that not all phrases will be indexed in Entrez nucleotide or protein indexes. Therefore, you may also wish to re-run the search without enclosing the terms in quotation marks. Carefully check your results to see if your phrase was found. In the "16S RNA" example, documents with the phrase '16S RNA' are retrieved, but documents with the phrase 16S RNA gene are not retrieved. PubMed handles quoted searches differently and for details on this, see the PubMed help document.

Compare the number of records retrieved in the Nucleotide database as of February 2006 in each case below. For example, the quoted phrase "16S RNA" retrieves fewer documents when compared with the search 16S AND RNA..

Example
16S rRNA retrieves 299,494
“16S rRNA” retrieves 85,500

Searching for Authors

Enter author names in the format: last name plus initials (e.g., johnson d). Do not use punctuation. This format instructs Entrez to search only the Author field. Entrez automatically truncates on the author's name to account for various initials and designations, such as Jr. or 2nd. If a last name is the only term entered in the query box (e.g., johnson), Entrez will search All Fields for that term. One can also use the Entrez Index term limiter : Johnson a[AUTHOR]. This also instructs Entrez to search only the author field.

Searching for Unique Identifiers

Unique identifiers can be **accession numbers**, which apply to a complete sequence record, or **sequence identification numbers**, which apply to the individual sequences within a record.

The format of **accession numbers** varies, depending upon the source database. (As noted above in The Databases section, each data domain in Entrez contains records from a number of different sources.) Some examples of typical accession number formats are below. The Sample GenBank Record [http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html] contains additional

detail about accession numbers [http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html#AccessionB]. Click here for a list of the Accession Number Prefixes [http://www.ncbi.nlm.nih.gov/Sequin/acc.html] .

Type of Record	Sample Accession Format
GenBank/EMBL/DDBJ Nucleotide Sequence Records	One letter followed by five digits, e.g.: U12345 Two letters followed by six digits, e.g.: AY123456, AF123456
GenPept Sequence Records (which contain the amino acid translations from GenBank/EMBL/DDBJ records that have a coding region feature annotated on them)	Three letters and five digits, e.g.: AAA12345
Protein Sequence Records from SWISS-PROT and PIR	All are six characters: Character/Format 1 [O,P,Q] 2 [0-9] 3 [A-Z,0-9] 4 [A-Z,0-9] 5 [A-Z,0-9] 6 [0-9] e.g.: P12345 and Q9JJS7
Protein Sequence Records from PRF	A series of digits (often six or seven) followed by a letter, e.g.: 1901178A
RefSeq [http://www.ncbi.nlm.nih.gov/RefSeq/key.html#accessions] Nucleotide Sequence Records	Two letters, an underscore bar, and six digits, e.g.: mRNA records (NM_*): NM_000492 genomic DNA contigs (NT_*): NT_000347 complete genome or chromosome (NC_*): NT_000907 genomic region (NG_*): NG000019
RefSeq [http://www.ncbi.nlm.nih.gov/RefSeq/key.html#accessions] Protein Sequence Records	Two letters (NP), an underscore bar, and six digits, e.g.: NP_000483
RefSeq [http://www.ncbi.nlm.nih.gov/RefSeq/key.html#accessions] Model (predicted) Sequence Records from the Human Genome annotation process	Two letters (XM, XP, or XR), an underscore bar, and six digits, e.g.: XM_000483
Protein Structure Records	PDB accessions generally contain one digit followed by three letters, e.g.: 1TUP MMDB ID numbers generally contain four digits, e.g.: 3973 The record for the Tumor Suppressor P53 Complexed With DNA can be retrieved by either number above.

There are two types of sequence identification numbers:

GI numbers:

a series of digits that are assigned consecutively by NCBI to each sequence it processes.

Version numbers:

consist of the accession number followed by a dot and a version number.

For example, the RefSeq [http://www.ncbi.nlm.nih.gov/RefSeq/key.html#accessions] record for the *Homo sapiens* cystic fibrosis transmembrane conductance regulator (cftr) has the accession number NM_000492. The record contains one nucleotide sequence and one amino acid translation, which have the following sequence identifers:

Nucleotide sequence:

GI: 6995995

VERSION: NM_000492.2

Protein translation:

GI: 6995996

VERSION: NP_000483.2

If a sequence changes in any way, it receives a new GI number, and the version number is incremented by one. The Sample GenBank Record [http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html] contains additional detail about GI [http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html#GInB] and Version [http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html#VersionB] sequence identification numbers.

Searching by Molecular Weight

NCBI implemented a "Molecular Weight" search field for searches of the Entrez Proteins database at the request of the mass spectrometry group at NIH. Dr. Lewis Pannell provided technical advice.

The Molecular Weight field can be queried as a single molecular weight:

2002 [Molecular Weight]

or a range of weights:

2002:2009 [Molecular Weight]

or either expression can be combined with other Entrez search terms, for example, to limit by organism:

002002:002009 [Molecular Weight] AND human [Organism]

The square brackets can contain the full spelling of the search field, as in the examples above, or the abbreviation [MOLWT] in upper- or lowercase.

Note also that where cleavage products are annotated with features, the molecular weight of each cleavage product is calculated, not the molecular weight of the whole protein. Thus, you may retrieve a large protein when querying with a small molecular weight; be sure to check the feature table of the protein record to see if it has cleavage products.

How the Molecular Weight is calculated:

If cleavage products are annotated, molecular weight is calculated for each cleavage product, not for the whole protein. Cleavage products are not consistently annotated, but we have done our best to detect the annotations across different database styles. For example, cleavage products are annotated as "matp" in GenBank but as "Region" with "/region_name=Mature chain" in SWISS-PROT.

Note that this means that more than one molecular weight may point to a single protein record!

If only a signal peptide is annotated, it is removed, and the molecular weight is calculated on the rest of the protein.

If there are no such features on the protein, then the molecular weight for the whole protein is calculated. In this case, a check is made for an initial Met, and it is not included in the calculation if found.

If completely unknown amino acids (e.g., "X") are found, a molecular weight is not calculated. Ambiguous amino acids are calculated as one of their possible forms:

B means D or N -- molecular weight is calculated as D

Z means E or Q -- molecular weight is calculated as E

Molecular weight is calculated as part of the indexing process for protein records in Entrez. Entrez's molecular weight is an average molecular weight, not monoisotopic. Masses are rounded to the nearest integer. The weights are present only in the Molecular Weight index and are not shown explicitly on the protein sequence records.

Range Searching

Range searching can be done on four data elements: accession numbers [ACCN], sequence length [SLEN], molecular weights [MOLWT], and dates [MDAT] and [PDAT]. The range operator is the colon (:), and the appropriate field qualifier should be included in square brackets after the second term. Field qualifiers are case insensitive, therefore either [ACCN] or [accn] will work. It is not necessary to include a space between the search term and the field qualifer, although that can be done, if desired.

Example searches:

Range of accession numbers:

AF114696:AF114714[ACCN]

Note: It is not possible to search for a range of sequence identification numbers (known as GI [http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html#GInB] numbers and Version [http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html#VersionB]) numbers.

Range of sequence lengths:

3000:4000[SLEN]

Range of molecular weights can be searched in the Protein database:

2002:2009[MOLWT]

Note: Additional information about Searching by Molecular Weight is included above.

A range search can also be combined with other Entrez search terms, for example, to limit by organism:

Select a protein database, enter this search in the query box: 2002:2009[MOLWT] AND human [ORGN]

In either the nucleotide or protein database: 3000:4000[SLEN] AND human[orgn]

To create a range, search for 11 through 999,999 nucleotide bases. Enter: 11:999999[SLEN]

Range of dates: 1998/02:2000/01/25[MDAT]

Truncating

Truncating search terms is a convenient way to find all the records that contain terms that begin with a given text string. Place an asterisk (*) at the end of a search term to find all records with a term that begins with that text string. For example, the truncated search term "immunoglob*" will retrieve all records in the database that contain the word immunoglobulin, immunoglobulins, immunoglobin, and immunoglobins. Searching Immunoglob* in the Entrez Protein database :

Example
immunoglob* results in 99,760 records

Entrez searches the first 600 variations of a truncated term. If a truncated term produces more than 600 variations, which is possible with terms like "bac*," Entrez gives the following warning:

" Wildcard search for 'bac*' used only the first 600 variations. Lengthen the root word to search for all endings."

Phrases that include a space in the word after the asterisk will NOT be retrieved. For example, if you search "chromo*," the documents retrieved will contain terms like chromobacterium but not chromo helicase.

Left-handed truncation is not possible (e.g., "*bacterium").

Combining Sets

Use your search History to combine documents retrieved with different search terms at different times during your search session. For example, search the Nucleotide database for HIV. This search retrieves over 188,000 documents. Now search the Nucleotide database for protease. This search retrieves over 92,000 documents. Now click on the History for the Nucleotide database.

The results for the HIV and protease search terms are saved as Search Sets #1 and #2, respectively. In the query box, type #1 AND #2 and select Go. This search combines the documents in Search Set #1 (HIV) with the documents in Search Set #2 (protease) and retrieves only those documents that are in both sets.

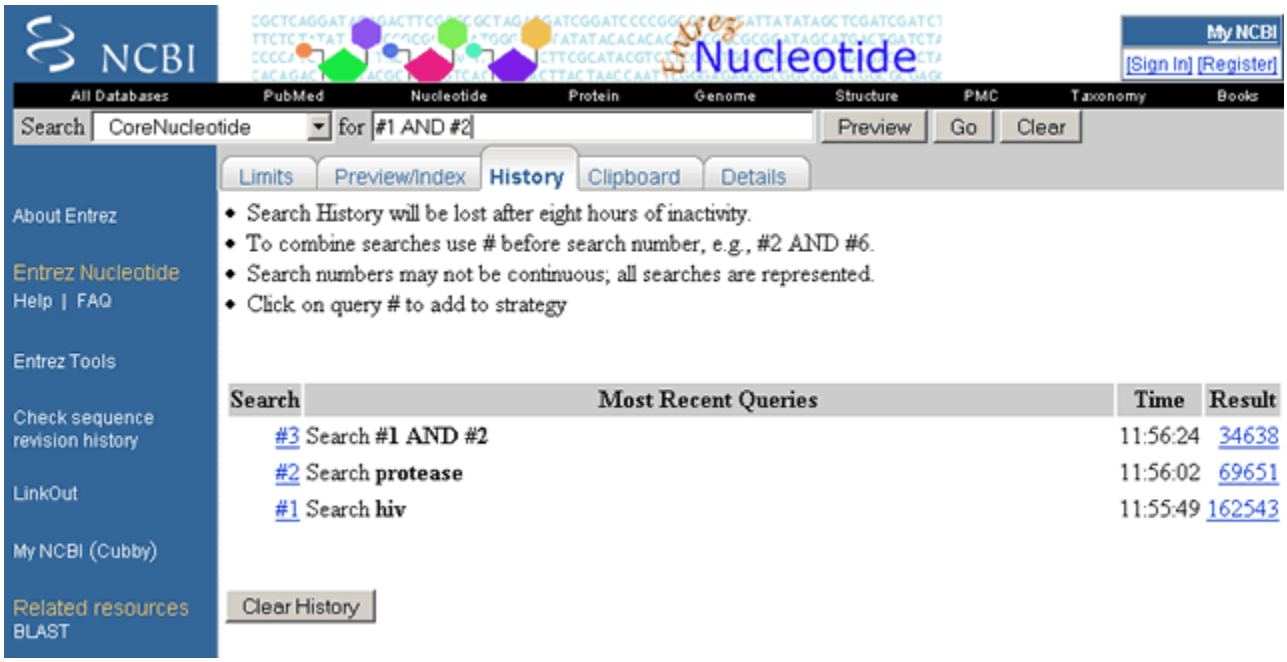
Click on History again and note that Search Set #3 = #1 AND #2.

Remember, this History is for the Nucleotide database only, and it will be lost after 8 hours of inactivity.

Search numbers may not be continuous; all searches are represented.

See Boolean Operators and Using Your History for more information and examples.

Sample Search History in CoreNucleotide to Combine Search Terms



Refining Your Search

Sometimes it is necessary to refine your search statement by using the Limits, Preview/Index, and History options of a given Entrez database. Review the search fields and Boolean Operators to effectivley use the Entrez Limits and Preview/Index options.

Limits

Limits allow restriction of a search to a defined subset of the database. Limits can be set to restrict a search to a particular database field (e.g., the Author field). Limits can be set to search everything but a particular type of data (e.g., “exclude patent records”). Alternatively, limits can be set to search only a particular type of data (e.g., Genomic RNA/DNA) or to search only data from a particular source database (e.g., EMBL). Date limits and sequence length limits are also possible.

The contents of each Entrez database differ, and therefore the Limits available for each database differ. See the “Limits Available by Database Summary” in the Summary Matrices section (see Table 1, Table 2 , Table 3 and Table 4) of this introduction. See also the Using Limits section of this document for help in using limits in your search.

Limits available for the Component Nucleotide and CoreNucleotide database

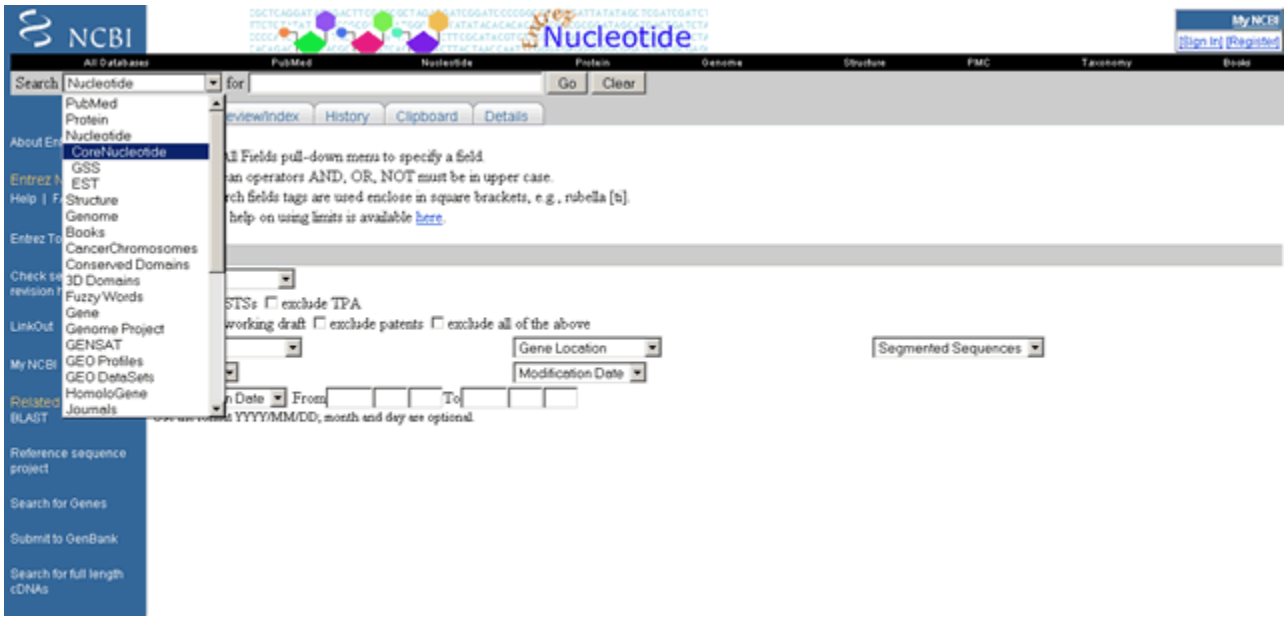


Table 1. Limits Available by Database

Databases								
Limits	Nucleotide	CoreNucleotide	EST	GSS	Protein	Genome	Structure	PopSet
Search Fields	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Exclude ESTs	Yes	No	No	No	No	No	No	No
Exclude STSs	Yes	Yes	No	No	No	No	No	No
Exclude GSSs	Yes	No	No	No	No	No	No	No
Exclude Working Draft	Yes	Yes	No	No	No	No	No	No
Exclude Patents	Yes	Yes	No	No	Yes	No	No	No
Molecule Type	Yes	Yes	Yes	Yes	No	No	No	No
Gene Location	Yes	Yes	No	No	Yes	No	No	No
Segmented Sequences	Yes	Yes	no	no	Yes	No	No	No
Database Source	Yes	Yes	Yes	Yes	Yes	No	No	No
Modification Date	Yes	Yes	Yes	Yes	Yes	No	No	No

Table 2. Search Fields Available by Database

Search Field Descriptions and Qualifiers	Nucleotide	Protein	Databases Genome	Structure	PopSet
Accession	Yes	Yes	Yes	Yes	Yes
All Fields	Yes	Yes	Yes	Yes	Yes
Author Name	Yes	Yes	Yes	Yes	Yes
EC/RN Number	Yes	Yes	Yes	Yes	Yes
Feature Key	Yes	No	Yes	No	Yes
Filter	Yes	Yes	Yes	Yes	Yes
Gene Name	Yes	Yes	Yes	No	Yes
Issue	Yes	Yes	Yes	Yes	Yes
Journal Name	Yes	Yes	Yes	Yes	Yes
Keyword	Yes	Yes	Yes	No	Yes
Modification Date	Yes	Yes	Yes	Yes	Yes
Molecular Weight	No	Yes	No	No	No
Organism	Yes	Yes	Yes	Yes	Yes
Page Number	Yes	Yes	Yes	Yes	Yes
Primary Accession	Yes	Yes	Yes	No	Yes
Properties	Yes	Yes	Yes	No	Yes
Protein Name	Yes	Yes	Yes	No	Yes
Publication Date	Yes	Yes	Yes	Yes	Yes
SeqID String	Yes	Yes	Yes	No	Yes
Sequence Length	Yes	Yes	Yes	No	No
Substance Name	Yes	Yes	No	Yes	No
Text Word	Yes	Yes	Yes	Yes	Yes
Title Word	Yes	Yes	Yes	No	No
Uid	No	No	No	No	No
Volume	Yes	Yes	Yes	Yes	Yes

Table 3. Search Field Descriptions and Qualifiers Corenucleotide Database

Nucleotide/CoreNucleotide Index Search Field	Definition	Qualifier
Accession	Contains the unique accession number of the sequence or record, assigned to the nucleotide, protein, structure, genome record, or PopSet by a sequence database builder. The Structure database accession index contains the PDB IDs but not the MMDB IDs. example : AF123456[accn]	[ACCN] or [ACCESSION]
All Fields	Contains all terms from all searchable database fields in the database.	[ALL] or [ALL FIELDS]
Author	Contains all authors from all references in the database records. The format is last name space first initial(s), without punctuation (e.g., marley jf).	[AUTH] or [AUTHOR]
EC/RN Number	Number assigned by the Enzyme Commission or Chemical Abstract Service (CAS) to designate a particular enzyme or chemical, respectively.	[ECNO]

Nucleotide/CoreNucleotide Index Search Field	Definition	Qualifier
Feature Key	Contains the biological features assigned or annotated to the nucleotide sequences and defined in the DDBJ/EMBL/GenBank Feature Table (http://www.ncbi.nlm.nih.gov/projects/collab/FT/index.html). Not available for the Protein or Structure databases.	[FKEY]
Filter	Contains predetermined or filtered subsets of the various databases. These subsets or filters are created by grouping records that are commonly linked to other Entrez databases or within the same database. For example, the PopSet database Filter index includes PopSet all, PopSet medline, PopSet nucleotide, and PopSet protein. The PopSet medline filter includes all PopSet records with links to PubMed; the PopSet nucleotide filter includes all PopSet records with links to the nucleotide database; and, the PopSet protein filter includes all PopSet records with links to the protein database. The PopSet all filter includes all PopSet records.	[FILT] or [SB]
Gene Name	Contains the standard and common names of genes found in the database records. This field is not available in Structure database.	[GENE]
Issue	Contains the issue number of the journal in which the data were published.	[ISS] or [ISSUE]
Keyword	Contains special index terms from the controlled vocabularies associated with the GenBank, EMBL, DDBJ, SWISS-Prot, PIR, PRF, or PDB databases. Browse the Keyword indexes of the individual databases to become familiar with these vocabularies. A Keyword index is not available in the Structure database.	[KYWD] or [KEYWORD]
Journal Name	Contains the name of the journal in which the data were published. Journal names are indexed in the database in abbreviated form (e.g., J Biol Chem). Journals are also indexed by their by ISSNs. Browse the index if you do not know the ISSN or are not sure how a particular journal name is abbreviated.	[JOUR] or [JOURNAL]
Modification Date	Contains the date that the most recent modification to that record is indexed in Entrez, in the format YYYY/MM/DD (e.g., 1999/08/05). A year alone, (e.g., 1999) will retrieve all records modified for that year; a year and month (e.g., 1999/03) retrieves all records modified for that month that are indexed in Entrez.	[MDAT]
Organism	Contains the scientific and common names for the organisms associated with protein and nucleotide sequences.	[ORGN] or [ORGANISM]

Nucleotide/CoreNucleotide Index Search Field	Definition	Qualifier
Page Number	Contains the number of the first journal page of the article in which the data were published.	[PAGE]
Primary Accession	Contains the primary accession number of the sequence or record, assigned to the nucleotide, protein, structure, genome record, or PopSet by a sequence database builder. A Primary Accession index is not available in the Structure database.	[PACC]
Properties	Contains properties of the nucleotide or protein sequence. For example, the Nucleotide database's Properties index includes molecule types, publication status, molecule locations, and GenBank divisions. A Properties index is not available in the Structure database.	[PROP]
Protein Name	Contains the standard names of proteins found in database records. Common names may not be indexed in this field so it is best to also consider All Fields or Text Words. A Protein Name index is not available in the Structure database.	[PROT]
Publication Date	Contains the date that records are released into Entrez, in the format YYYY/MM/DD (e.g., 1999/08/05). It is the date the entry first appeared in GenBank explicitly indexed in Entrez. A year alone, (e.g., 1999) will retrieve all records for that year; a year and month (e.g., 1999/03) will retrieve all records released into GenBank for that month.	[PDAT]
SeqID String	Contains the special string identifier, similar to a FASTA identifier, for a given sequence. A SeqID String index is not available in the Structure database.	[SQID]
Sequence Length	Contains the total length of the sequence. Sequence Length indexes are not available in the Structure or PopSet databases.	[SLEN]
Substance Name	Contains the names of any chemicals associated with this record from the CAS registry and the MEDLINE Name of Substance field. Substance Name indexes are not available in the Genome or PopSet databases.	[SUBS]
Text Word	Contains all of the "free text" associated with a record	[WORD]
Title	Includes only those words found in the definition line of a record. The definition line summarizes the biology of the sequence and is carefully constructed by database staff. A standard definition line will include the organism, product name, gene symbol, molecule type and whether	[TITL]

Nucleotide/CoreNucleotide Index Search Field	Definition	Qualifier
	it is a partial or complete cds. Title Word indexes are not available in the Structure or PopSet databases.	
Volume	Contains the volume number of the journal in which the data were published.	[VOL]

Table 4. Search Field Descriptions and Qualifiers Protein Database

Protein Database Index Search Field	Definition	Qualifier
Accession	Contains the unique accession number of the sequence or record, assigned to the nucleotide, protein, structure, genome record, or PopSet by a sequence database builder. The Structure database accession index contains the PDB IDs but not the MMDB IDs.	[ACCESSION] or [ACCN]
All Fields	Contains all terms from all searchable database fields in the database.	[ALL] or [ALL FIELDS]
Author	Contains all authors from all references in the database records. The format is last name space first initial(s), without punctuation (e.g., marley jf).	[AUTH] or [AUTHOR]
EC/RN Number	Number assigned by the Enzyme Commission or Chemical Abstract Service (CAS) to designate a particular enzyme or chemical, respectively.	[ECNO]
Filter	Contains predetermined or filtered subsets of the various databases. These subsets or filters are created by grouping records that are commonly linked to other Entrez databases or within the same database. For example, the PopSet database Filter index includes PopSet all, PopSet medline, PopSet nucleotide, and PopSet protein. The PopSet medline filter includes all PopSet records with links to PubMed; the PopSet nucleotide filter includes all PopSet records with links to the nucleotide database; and, the PopSet protein filter includes all PopSet records with links to the protein database. The PopSet all filter includes all PopSet records.	[FILT] or [SB] or [FILTER]
Gene Name	Contains the standard and common names of genes found in the database records. This field is not available in Structure database.	[GENE]
Issue	Contains the issue number of the journal in which the data were published.	[ISS] or [ISSUE]
Keyword	Contains special index terms from the controlled vocabularies associated with the GenBank, EMBL, DDBJ, SWISS-	[KYWD] or [KEYWORD]

Protein Database Index Search Field	Definition	Qualifier
	Prot, PIR, PRF, or PDB databases. Browse the Keyword indexes of the individual databases to become familiar with these vocabularies. A Keyword index is not available in the Structure database.	
Journal	Contains the name of the journal in which the data were published. Journal names are indexed in the database in abbreviated form (e.g., J Biol Chem). Journals are also indexed by their by ISSNs. Browse the index if you do not know the ISSN or are not sure how a particular journal name is abbreviated.	[JOUR] or [JOURNAL]
Modification Date	Contains the date that the most recent modification to that record is indexed in Entrez, in the format YYYY/MM/DD (e.g., 1999/08/05). A year alone, (e.g., 1999) will retrieve all records modified for that year; a year and month (e.g., 1999/03) retrieves all records modified for that month that are indexed in Entrez.	[MDAT]
Molecular Weight	Molecular weight of a protein, in Daltons (Da), calculated by the method described in the Searching by Molecular Weight section of the Entrez help document. Note that molecular weight must be entered as a fixed 6 digit field, filled with leading zeros (not letter O), e.g., 002002 [MOLWT]	[MOLWT]
Organism	Contains the scientific and common names for the organisms associated with protein and nucleotide sequences.	[ORGN] or [ORGANISM]
Page Number	Contains the number of the first journal page of the article in which the data were published.	[PAGE]
Primary Accession	Contains the primary accession number of the sequence or record, assigned to the nucleotide, protein, structure, genome record, or PopSet by a sequence database builder. A Primary Accession index is not available in the Structure database.	[PACC]
Properties	Contains properties of the nucleotide or protein sequence. For example, the Nucleotide database's Properties index includes molecule types, publication status, molecule locations, and GenBank divisions. A Properties index is not available in the Structure database.	[PROP] or [PROPERTIES]
Protein Name	Contains the standard names of proteins found in database records. Common names may not be indexed in this field so it is best to also consider All Fields or Text Words. A Protein Name index is not available in the Structure database.	[PROT] or [PROTEIN NAME]

Protein Database Index Search Field	Definition	Qualifier
Publication Date	Contains the date that records are released into Entrez, in the format YYYY/MM/DD (e.g., 1999/08/05). It is the date the entry first appeared in GenBank explicitly indexed in Entrez. A year alone, (e.g., 1999) will retrieve all records for that year; a year and month (e.g., 1999/03) will retrieve all records released into GenBank for that month.	[PDAT] or [PUBLICATION DATE]
SeqID String	Contains the special string identifier, similar to a FASTA identifier, for a given sequence. A SeqID String index is not available in the Structure database.	[SQID] or [SEQUID STRING]
Sequence Length	Contains the total length of the sequence. Sequence Length indexes are not available in the Structure or PopSet databases.	[SLEN] or [SEQUENCE LENGTH]
Substance Name	Contains the names of any chemicals associated with this record from the CAS registry and the MEDLINE Name of Substance field. Substance Name indexes are not available in the Genome or PopSet databases.	[SUBS] or [Substance Name]
Text Word	Contains all of the "free text" associated with a record	[WORD] or [Text Word]
Title	Includes only those words found in the definition line of a record. The definition line summarizes the biology of the sequence and is carefully constructed by database staff. A standard definition line will include the organism, product name, gene symbol, molecule type and whether it is a partial or complete cds. Title Word indexes are not available in the Structure or PopSet databases.	[VOL] or [VOLUME]
Volume	Contains the volume number of the journal in which the data were published.	[VOL] or [VOLUME]

Using Limits

Limits are used to refine search results to retrieve only the most relevant documents. In other words limits remove unneeded or unwanted documents. This section provides examples for using limits to:

- Limit a Search to a Particular Database Field
- Exclude Certain Kinds of Sequences
- Limit the Search to a Particular Molecule Type
- Limit the Search to a Particular Gene Location

- Display Only the Master or Only the Parts of Segmented Sets of Sequences
- Limit the Search to Records from a Particular Sequence Database
- Limit the Search by Date
- Using More Than One Limit at a Time

See the summary matrices tables: Table 1, Table 2, Table 3, and Table 4 to review the limits available for each database.

Limit a Search to a Particular Database Field

Example: You are only interested in nucleotide sequences from the mouse:

1. Starting at the NCBI Home Page www.ncbi.nlm.nih.gov, select the Nucleotide database from the Search pull-down menu and select Go to get to the Nucleotide database page.
2. Select the tab entitled Limits.
3. In the "Limited To:" section, select Organism from the Search Field pull-down menu.
4. Type "mouse" without quotes in the query box and select Go.

On the results screen, note that the check box next to Limits is checked, indicating that Limits are selected and active. Beneath the check box, the selected and active limits are highlighted in yellow (i.e., Field: Organism). You may also link to the results in CoreNucleotide, EST, or GSS.

Note: Always remember to uncheck the box in the Limits tab when you have completed your limits search, as the Limits you have selected will remain active for future searches.

Example: You are only interested in protein sequences that are fewer than 50 amino acids in length:

1. Starting at the NCBI Homepage www.ncbi.nlm.nih.gov, select the Protein database from the Search pull-down menu and select Go to get to the Protein database page.
2. Select Limits.
3. In the "Limited To:" section, select Sequence Length from the Search Field pull-down menu.
4. Type "0:50" without quotes in the query box and select Go.

On the results screen, note that the check box next to Limits is checked, indicating that Limits are selected and active. Beneath the check box, the selected and active limits are highlighted in yellow (i.e., Field: Sequence Length).

Note: Always remember to uncheck the box in the Limits tab when you have completed your limits search, as the Limits you have selected will remain active for future searches.

Exclude Certain Kinds of Sequences

Example: You are interested in mitochondrial carriers, but you do not want any patent sequences:

1. Select the CoreNucleotide database from the Search pull-down menu.
2. Type "mitochondrial carrier" without quotes in the query box.
3. In the "Limited To:" section, check the box next to "exclude Patents" and select Go.

On the results screen, note that the check box next to Limits is checked, indicating that Limits are selected and active. Beneath the check box, the selected and active limits are highlighted in yellow (i.e., Limits: exclude Patents).

In the Nucleotide database, you can exclude EST, STS, GSS, working drafts, TPA [http://www.ncbi.nih.gov/Genbank/TPA.html] and/or Patent sequences. In the CoreNucleotide database, you can exclude STS, working drafts, TPA, and/or patent sequences. In the Protein database, you can exclude TPA and Patent sequences.

Limit the Search to a Particular Molecule Type

Example: You are only interested in Cryptosporidium ribosomal RNA sequences:

1. Starting at the NCBI Homepage www.ncbi.nlm.nih.gov, select the Nucleotide database from the Search pull-down menu and select Go to get to the Nucleotide database page.
2. Type "cryptosporidium" without quotes in the query box.
3. Select Limits.
4. In the "Limited To:" section, select the "Molecule" pull-down menu and choose rRNA and select Go.

On the results screen, note that the check box next to Limits is checked, indicating that Limits are selected and active. Beneath the check box, the selected and active limits are highlighted in yellow (i.e., Limits: rRNA). You may also link to the results in CoreNucleotide, EST, or GSS.

Limit the Search to a Particular Gene Location

Example: You are interested in the genes in the chloroplast of flowering plants:

1. Select the Nucleotide database from the black menu bar or the Search pull-down menu.
2. Type "flowering plants" without quotes in the query box.
3. Select Limits.
4. In the "Limited To:" section, select the "Gene Location" pull-down menu and choose Chloroplast and select Go.

On the results screen, note that the check box next to Limits is checked, indicating that Limits are selected and active. Beneath the check box, the selected and active limits are highlighted in yellow (i.e., Limits: Chloroplast). You may also link to a majority of the records in CoreNucleotide.

Display Only the Master or Only the Parts of Segmented Sets of Sequences

Example: You are interested in the cystic fibrosis transmembrane conductance regulator (CFTR) gene. You know that there are several segmented sets of sequences associated with the CFTR gene. But you are only interested in displaying the master record of any segmented sets associated with the CFTR gene:

1. Select the Nucleotide database from the black menu bar or the Search pull-down menu.
2. Type "cftr" without quotes in the query box.
3. Select Limits.
4. In the "Limited To:" section, select the "Segmented Sequences" pull- down menu and choose "Show only master of set" and select Go.

On the results screen, note that the check box next to Limits is checked, indicating that Limits are selected and active. Beneath the check box, the selected and active limits are highlighted in yellow (i.e., Limits: Show only master of set).

Please note that this option does not allow you to limit the documents retrieved to only those containing segmented sequences. It simply allows you to control how segmented sets of sequences are displayed.

Limit the Search to Records from a Particular Sequence Database

Example: You are interested only in cysteine phosphatase protein sequences submitted directly to PIR:

1. Select the Protein database from the black menu bar or the Search pull-down menu.
2. Type "cysteine phosphatase" without quotes in the query box.
3. Select Limits.
4. In the "Limited To:" section, select the "Only from" pull-down menu and choose PIR and select Go.

On the results screen, note that the check box next to Limits is checked, indicating that Limits are selected and active. Beneath the check box, the selected and active limits are highlighted in yellow (i.e., Limits: PIR).

Limit the Search by Date

Example: You want to see any nucleotide sequences from pigs added to the database (or updated) in the last 30 days:

1. Select the Nucleotide database from the black menu bar or the Search pull-down menu.
2. Type "pigs" without quotes in the query box.
3. Select Limits.
4. In the "Limited To:" section, select Organism from the Search Field pull-down menu.
5. In the "Limited To:" section, select the "Modification Date" pull-down menu, choose "30 Days", and select Go.

On the results screen, note that the check box next to Limits is checked, indicating that Limits are selected and active. Beneath the check box, the selected and active limits are highlighted in yellow (i.e., Field: Organism, Limits: 30 Days).

Example: You want to retrieve all mouse or human protein sequences added to the database (or updated) during 1997:

1. Select the Protein database from the black menu bar or the Search pull-down menu.
2. Select Limits.
3. Type "mouse OR human" without quotes in the query box.
4. Select Limits.
5. In the "Limited To:" section, select Organism from the Search Field pull-down menu.
6. In the "Limited To:" section, select the "Modification Date" pull-down menu, and choose Modification Date (as opposed to Publication Date). In the date boxes, type the dates in the format YYYY/MM/DD. You can tab from box to box in the date fields. The From date is 1997/01/01, and the To date is 1997/12/31. Select Go.

On the results screen, note that the check box next to Limits is checked, indicating that Limits are selected and active. Beneath the check box, the selected and active limits are highlighted in yellow (i.e., Field: Organism, Limits: Modification Date, from 1997/01/01 to 1997/12/31).

Using More Than One Limit at a Time

As shown in the last two examples, you can use more than one limit at a time. Here is one more example using multiple limit features in an Entrez search.

Example: You are interested in the protein translations of human GenBank nucleotide sequences added to the protein database (or updated) in the last 30 days. You do not want patent records:

1. Select the Protein database from the black menu bar or the Search pull-down menu.
2. Select Limits.
3. Type "human" without quotes in the query box.
4. Select Limits.
5. In the "Limited To:" section, select Organism from the Search Field pull-down menu.
6. On the same screen, select the "exclude patents" check box, select GenBank from the "Only from" pull-down menu, and finally select "30 Days" from the Modification Date pull-down menu and select Go.

On the results screen, note that the check box next to Limits is checked, indicating that Limits are selected and active. Beneath the check box, the selected and active limits are highlighted in yellow (i.e., Field: Organism, Limits: Exclude patents, 30 Days, GenBank).

Using the Indexes

Indexes are used to browse and/or select the terms by which records and/or data are described. This section provides examples for using indexes to:

- Examine Search Field Indexes
- Browse, Select, and Search Terms
- Select, Combine, and Search Multiple Terms
- Select, Combine, and Search Multiple Terms from Multiple Indexes

Preview/index

Indexes are alphabetical lists of terms from searchable database fields. When indexes are displayed, they provide a way to browse the terms by which records and/or data are described. Entrez not only lets you browse indexes, you can also select terms to search directly from them.

As with limits, the indexes available for a particular database are dependent on the searchable fields of that database. See the “Indexes Available by Database” in tables for the summary matrices: Table 1, Table 2, Table 3 and Table 4.

The view below displays the entries listed alphabetically under "bacter" in the Organism index of the Nucleotide database. Specific indexes are selected from the "Add Term(s) to Query or View Index" pull-down menu. Search by typing search terms in the query box and select the Index button. Browse the terms by selecting the Up and Down buttons to scroll. See the Using the Indexes section of this document for help in using indexes in your search.

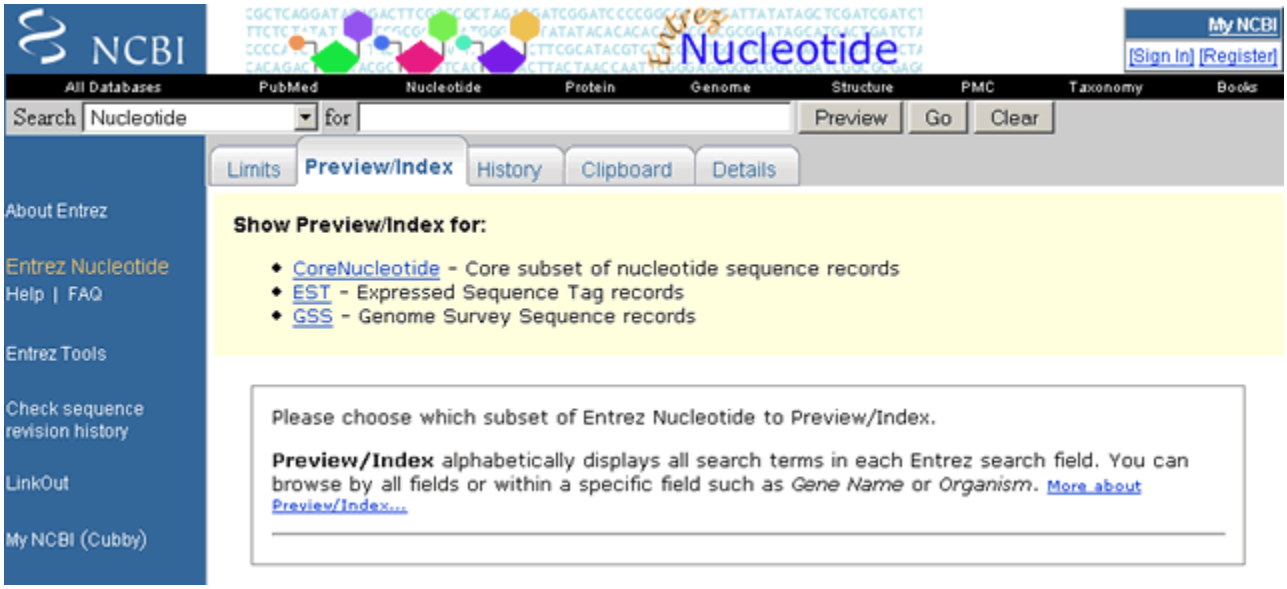
Component Nucleotide Databases All Fields Index

Available indexes for the Nucleotide database are shown here in the Nucleotide "Add Terms to Query or View Index" pull-down menu

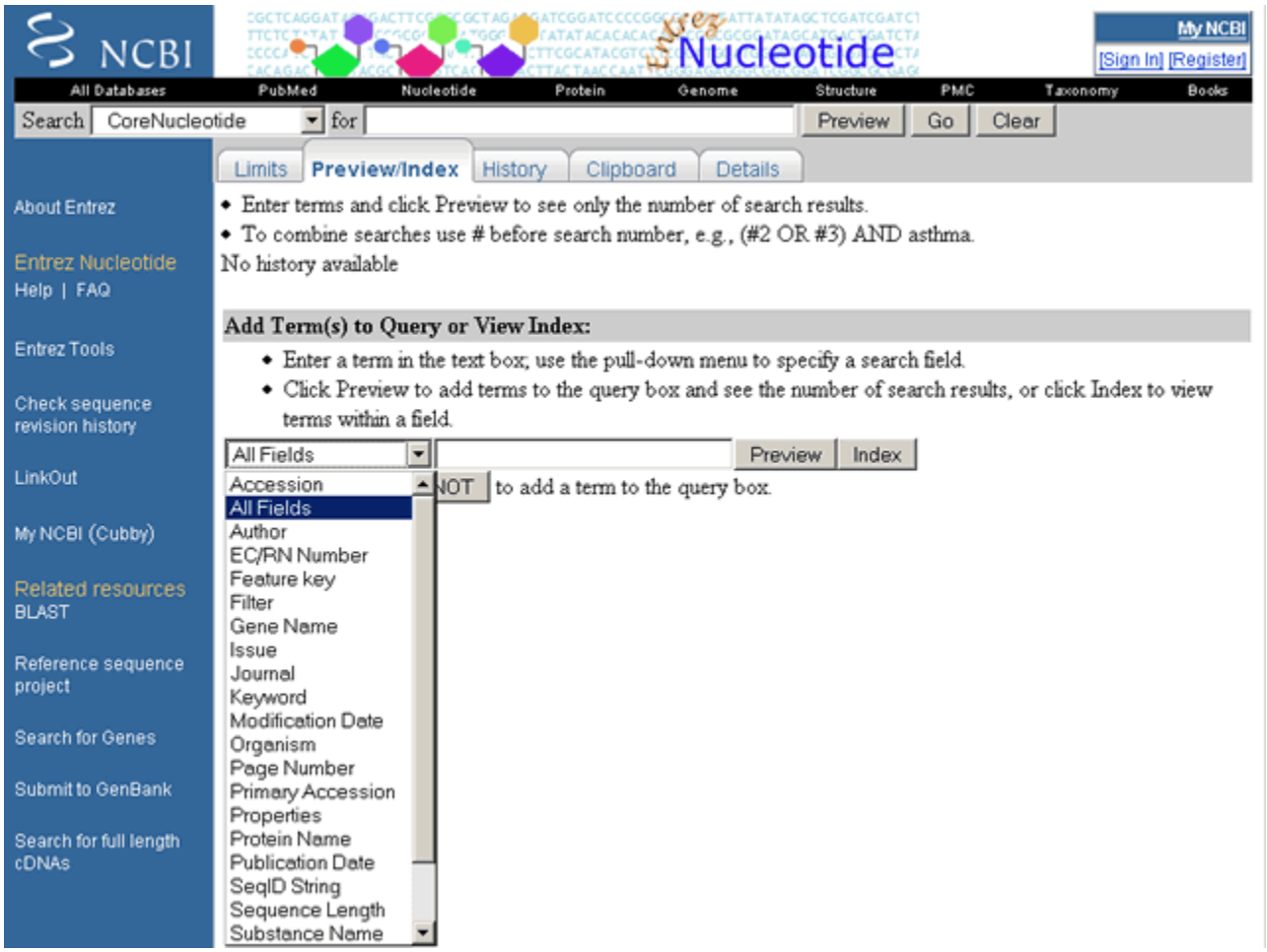
Available indexes for the Nucleotide database are shown below.

1. from the Nucleotide page, click on the 'Preview/Index' tab
2. add terms to 'Query' or 'View Index' section, click on the pulldown menu to get a list of index fields (see image below)

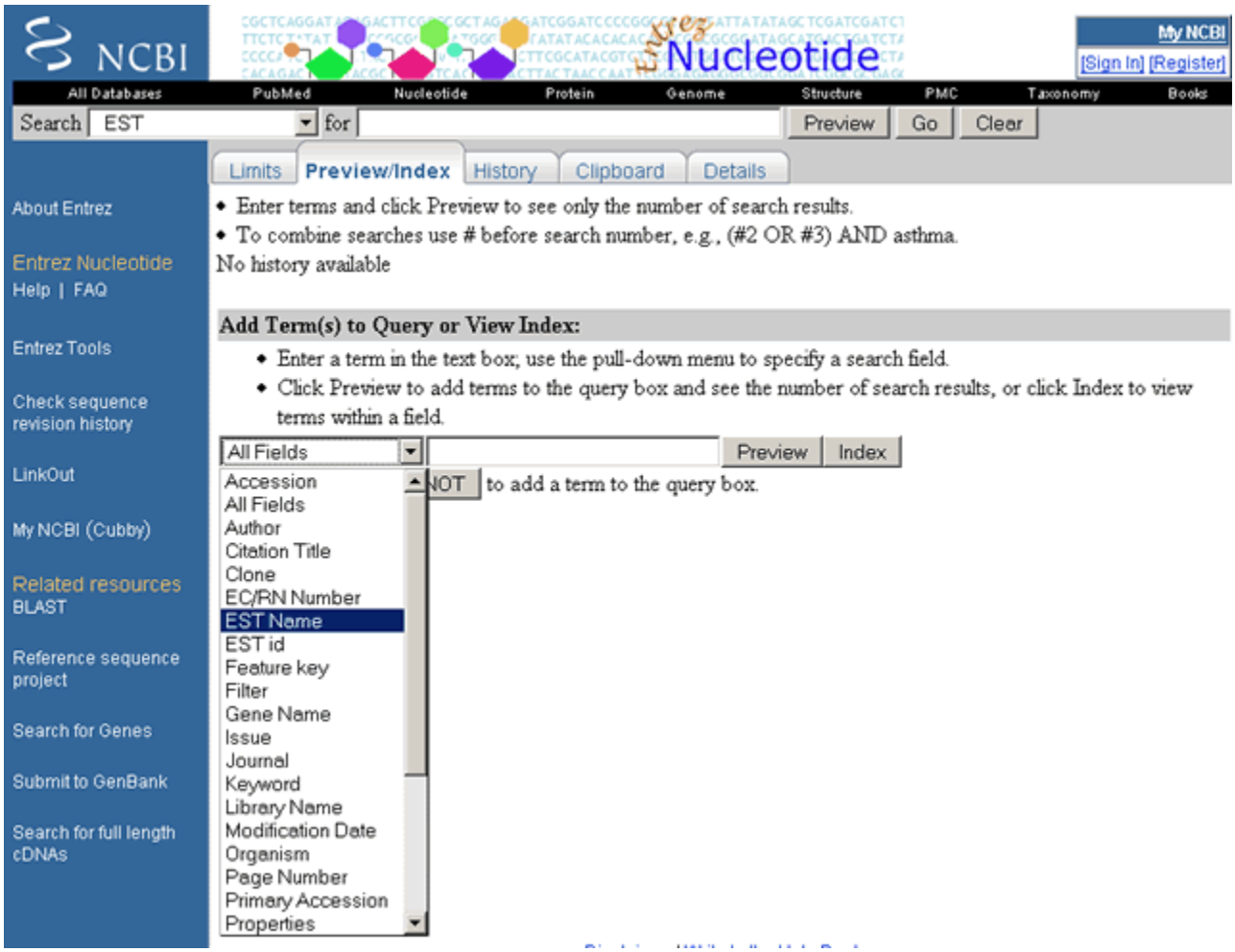
The Umbrella Nucleotide "Add Term(s) to Query or View Index" pull-down menu is shown



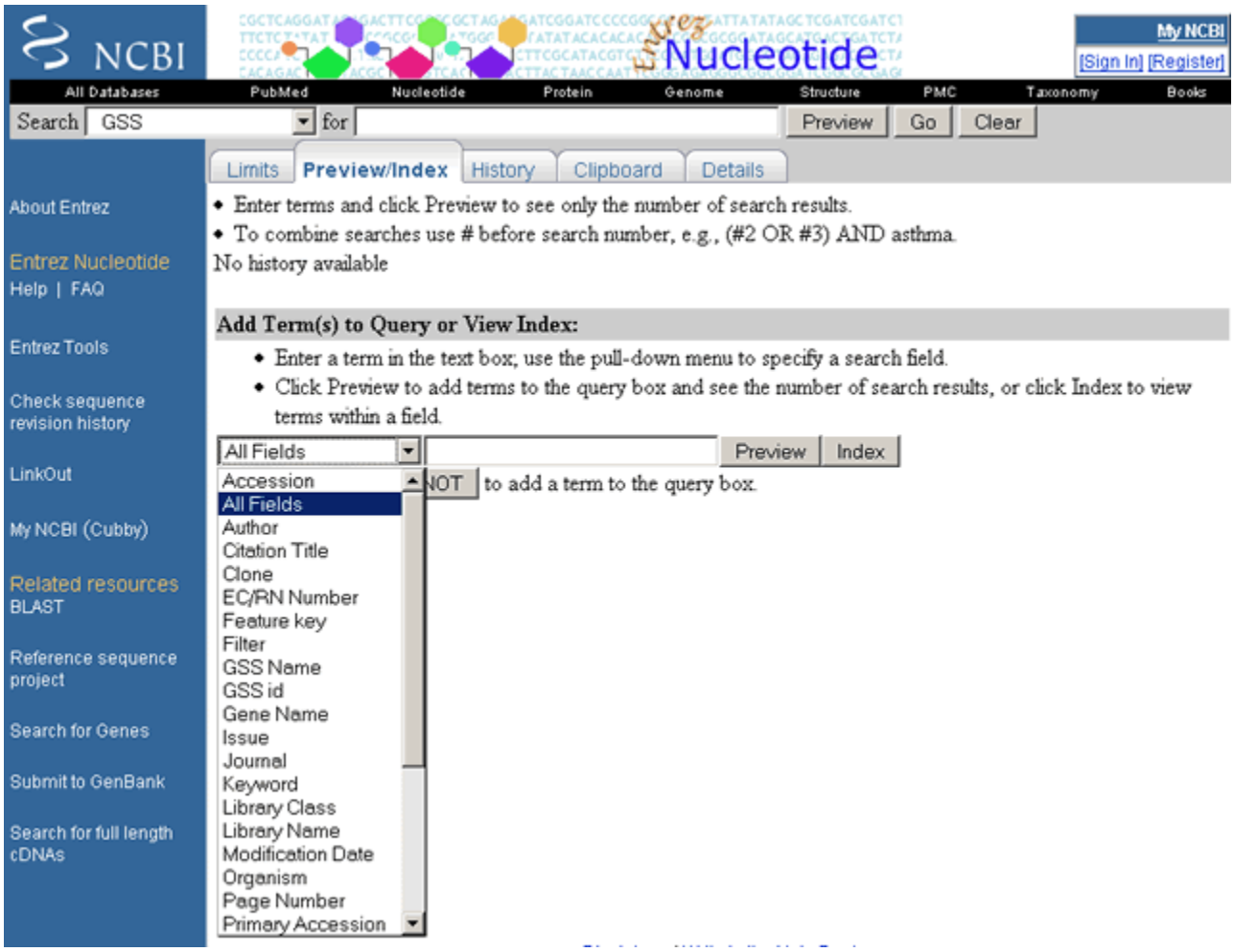
The CoreNucleotide "Add Term(s) to Query or View Index" pull-down menu is shown



The EST Nucleotide "Add Term(s) to Query or View Index" pull-down menu is shown



The GSS Nucleotide "Add Term(s) to Query or View Index" pull-down menu is shown



Examine Search Field indexes

Example: Examine the kind of information indexed in the Properties index of the Nucleotide database:

1. Select the Nucleotide database.
2. Select Preview/Index.
3. Select the Properties index from the Preview/Index pull-down menu.
4. Leave the Preview/Index search box empty and select the 'Index' button to see an alphabetical listing of all index terms within the Properties index.

Index entries are listed alphabetically and Entrez will begin the index display at the very first entry (i.e., biomol genomic).

Use the scroll bar to view more entries. Use the Down and Up buttons to display the next set of entries in either direction. The Properties search field and its corresponding index are very useful. This field contains information about the GenBank division to which the record belongs (i.e., gbdiv inv). It also describes the molecule type and location. The Properties field also describes such things as whether the sequence is part of a population study or segmented set.

Compare the Properties index of the Nucleotide database to the Properties index of the other databases. A Properties index is not available for the Structure database.

Example: Examine the kind of information indexed by the Feature key index of the Genome database.

1. Select the Genome database.
2. Select Index.
3. Select the Feature key index from the View Index pull-down menu.
4. Type "0" (the number zero) without quotes in the View Index query box and select View.

Use the scroll bar to view the entries. Use the Up and Down buttons to display the next set of entries in either direction. The Feature key search field and its corresponding index are also very useful. This field contains information about the biological features of the nucleotide sequences as annotated by submitters and database staff.

Browse, Select, and Search Terms

Example: You want to search all sequences in the GenBank EST division.

Simply select EST database and enter your search terms in the search box.

Select, Combine, and Search Multiple Terms

Example: You want all of the population sets for humans, mice, and Drosophila:

1. Select the PopSet database.
2. Select Index.
3. Select the Organism index from the View Index pull-down menu.
4. Type "human" without quotes in the View Index query box and select View.
5. View the list of entries and locate the "human" entry.
6. Select the "human" entry by clicking on it once.
7. Select the "human" entry as a search term by clicking "AND". Note that the term is now located in the Search query box as "human" [Organism].
8. Type "mouse" without quotes in the View Index query box and select View.
9. View the list of entries and locate the "mouse" entry.
10. Select the "mouse" entry by clicking on it once.
11. Select the "mouse" entry as a search term by clicking "OR". Note that the term is now located in the Search query box with the human term (i.e., "human"[Organism] OR "mouse"[Organism]).
12. Repeat steps 8-11 above for Drosophila so that the final search statement in the query box is:

"human"[Organism] OR "mouse"[Organism] OR "drosophila"[Organism]

13. Select Go to execute this search.

Select, Combine, and Search Multiple Terms from Multiple Indexes

Example: You want all protein kinase sequences from pigs:

1. Select the Protein database.
2. Select Index.
3. Select the Organism index from the View Index pull-down menu.
4. Type "pig" without quotes in the View Index query box and select View.
5. View the list of entries and locate the "pig" entry.

6. Select the "pig" entry by clicking on it once.

7. Select the "pig" entry as a search term by clicking "AND." Note that the term is now located in the Search query box as "pig" [Organism].

8. Select the Text Word index from the View Index pull-down menu.

9. Type "kinase" without quotes in the View Index query box and select View.

10. View the list of entries and locate the "kinase" entry.

11. Select the "kinase" entry by clicking on it once.

12. Select the "kinase" entry as a search term by clicking "AND". Note that the term is now located in the Search query box as "kinase" [Text Word] and that the final search statement in the query box is:

"pig"[Organism] AND "kinase"[Text Word]

13. Select Go to execute this search.

REMEMBER that Entrez processes complex search statements using Boolean Operators in a specific order as described in the Boolean Operators section above. You can always check the Details button to see how your final search statements were executed.

Using Your History

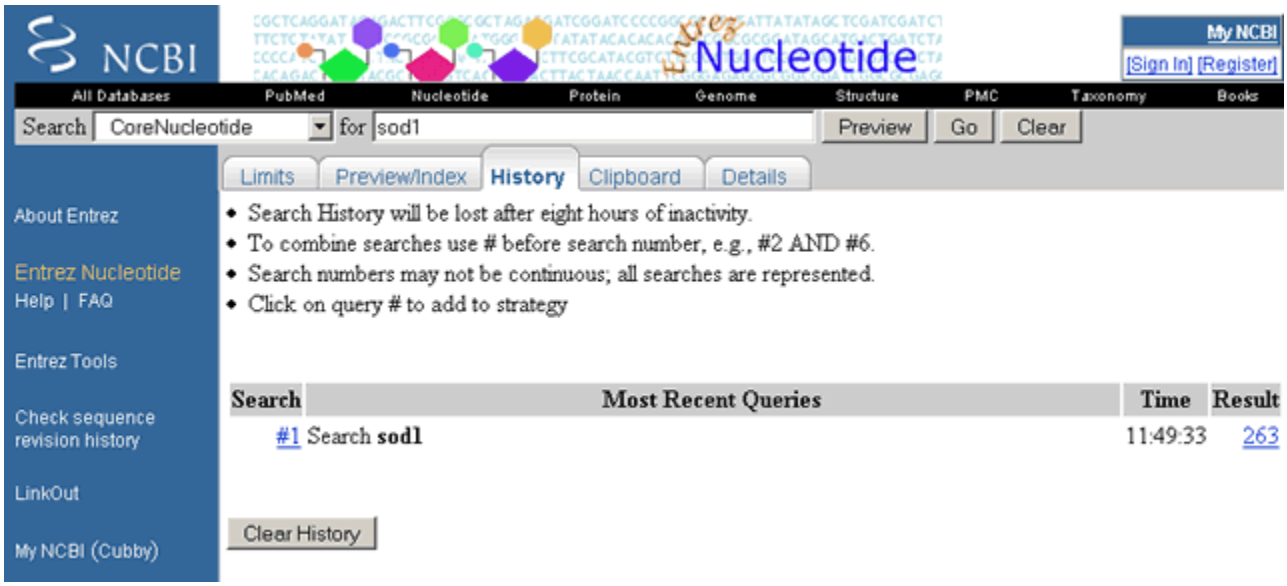
History provides a record of the searches performed during a search session. This section provides examples for using your search history.

History

Using the Preview option of Preview/Index allows a searcher to display the last three results for consecutive searches. A searcher can view the effect of each successive limit added to the search strategy. See the explanation of History in the next section for an option to see all search history for individual Entrez databases.

History provides a record of the searches performed during a search session. Histories are database specific. Each time search terms are typed into the query box and the search is executed, the search terms, the time the search was executed, and the search results are numbered consecutively and saved automatically in the History for that database. The History can be recalled at any time during a search session, but histories are lost after eight hours of inactivity. Use Histories to review, revise, or combine the results of earlier searches. See the Using Your History section of this document for help in using your search history.

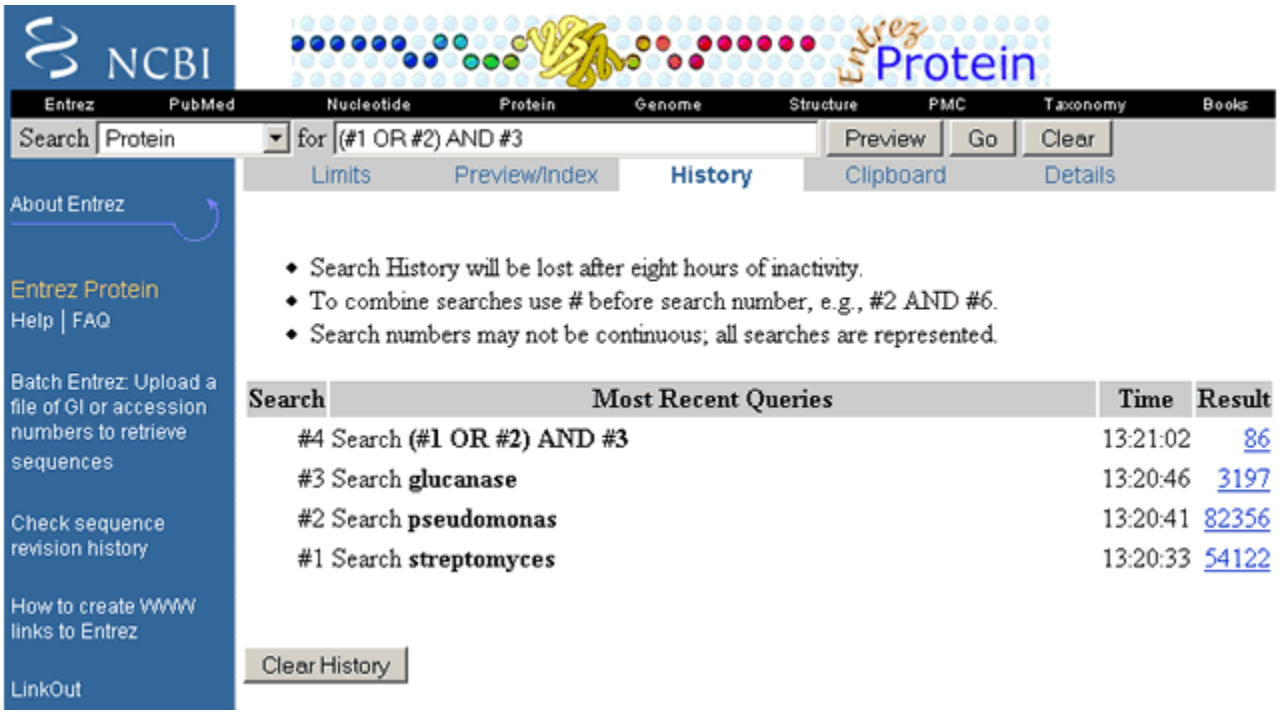
Sample History page for CoreNucleotide



Review a Search Session and Combine Results

Example: Search for Streptomyces, Pseudomonas, and glucanase and then use History to combine results:

1. Select the Protein database.
2. Type "streptomyces" in the query box and select Go.
3. Select Clear.
4. Type "pseudomonas" in the query box and select Go.
5. Select Clear.
6. Type "glucanase" in the query box and select Go.
7. Select History.
8. Review your search History and results. Note that each search statement is numbered. Also note the time and number of results for each search statement.
9. Combine the results of your earlier searches using the search numbers and Boolean operators. For example: ((#15) OR #16) AND #17. Select Go. Note: You can click on a "#" search result in 'History' to reveal an 'Options' menu.
10. Select History to once again review your search History and results.



Although search Histories are database specific, the History numbering system is continuous across all databases searched during a single search session. For instance, let us say you just finished searching the Protein database using the example above. Next you want to search the Structure database for similar information. You cannot use your Protein database search History in the Structure database. However, as you start searching the Structure database, Entrez sequentially numbers the search sets based on the last search query executed in any database. Therefore, in this example, the first search query executed in the Structure database is numbered search #30. The next search query executed is numbered search #31 and so on. Entrez will save a maximum of 100 queries at a time.

A final note on search histories. If you search the same query in the same database during the same search session, the search set will only be saved in the History one time. To perform batch searches in Entrez, see or select Entrez Tools.

Refine Search Results

Example: You are interested in any DNA sequences of the mouse fas antigen:

1. Select the Nucleotide database.

2. Type mouse[orgn] AND "fas antigen" with quotes around fas antigen in the query box and select Go.

3. The search retrieves over 30 documents. You do not want to review all the documents and decide you are really interested in any sequences with annotated exons or introns.

4. Select the 'History' tab.

5. Refine the results of your search using the search number and Boolean operators. For example: #1 AND (exon OR intron). Select 'Preview'.

6. Select 'History' to once again review your search History and results. Refining the search has reduced the number of retrieved documents to 9.

Writing Advanced Search Statements

Complex search statements can be written and executed directly from the the query box of any of the databases, as long as you obey some simple rules and use the correct syntax.

Perform a search by specifying the search terms, their fields, and the Boolean operations to perform on the term. Use the following syntax:

term [field] OPERATOR term [field]

Where term(s) are the search terms, the field(s) are the Search Field qualifiers from Table 1, Table 2, Table 3 , and Table 4 and the OPERATOR(s) are the Boolean Operators. Remember that Boolean operators are normally processed from left to right. If you want part of your Boolean expression to be processed out of order, enclose it in parentheses.

Example: Find all human nucleotide sequences with D-loop annotations.

In the Nucleotide database, use the following expression:

D-loop[FKEY] AND human[ORGN]

Example: Find all human protein sequences with lengths between 50 and 60 amino acids that were entered into the database during 1999.

In the Protein database, use the following expression:

human[ORGN] AND 50[SLEN]:60[SLEN] AND 1999[MDAT]

Example: Find Drosophila population studies published in the Journal of Molecular Evolution

In the PopSet database, use the following expression:

j mol evol[JOUR] AND drosophila[ORGN]

Displaying and Saving Results

Entrez displays search results as shown below:

The Search query box provides a summary of the database searched and the search terms as entered (i.e., "Search Nucleotide for hiv protease").

Clicking on the 'Links', adjacent each docsum, provides a dropdown list of databases that are linked to the individual document summary. Clicking on the 'Reports' button provides a dropdown list of data formats. The Sequence Revision History for the specific record can be viewed from the link 'Revision History'.

Document Summaries, or "docsums", are displayed for the "hiv protease" search within the Nucleotide Database.

Display, Show, Sort, and ‘Send to’ Options

Display - The default display format is the Summary format shown in the example above.

To change the Display format, select an alternate format from the format pull-down menu (i.e., Summary). The format automatically displays. See Table 5 for formats available by database for Nucleotide, CoreNucleotide, EST,GSS, PopSet, Protein, Genome, and Genome Project databases.

For a very large Nucleotide record, such as a genome or contig record, retrieve the document summary (docsum) for NT_037704, for example, you should not attempt to display in your browser, the full GenBank report with all the features and sequence, which is GenBank(Full) in the Display pull-down menu. Instructions for downloading or saving to file are below. Care should be excerized when attempting to download a very large record such as NC_000001 which has many features and a sequence of 247249719 bases. It saves to over 300MB. Opening such a large file requires a robust text editor and adequate space on your computer.

1) Select GenBank from the Display menu or click the NT_037704 link.

2) Select 'Send to File' and then, select the Cancel option on the save dialog box if it appears.

3) Select GenBank(Full) from the Display pull-down menu.

4) The save dialog box will appear showing the file name : "sequences.gbwithparts". Enter your desired directory/file location and click the Save option.

5) The entire GenBank record, including all of its features and annotations will be saved on your computer.

To view the "graphical view", click on the accession number to display the GenBank report format. Select 'Graphics' from the Display menu. The Entrez graphical view will appear.

Show - The default number of documents displayed is 20. The total number of pages is displayed on the far right of the results page

To change the number of documents displayed per page, select an alternate number from the pull-down menu (e.g., 50) and, once selected, it will automatically display the new number selected.

In the Nucleotide database, for example, results can be sorted by accession number. Select the 'Sort by' pulldown menu and select the accession option. Send to - provides an option menu for sending the results to either text, file, or clipboard.

See the display formats in Table 5. for a summary of the display formats available for the Nucleotide, CoreNucleotide, EST, GSS, Protein, PopSet, Genome, and Genome Project databases.

Table 5. Display formats for Nucleotide, CoreNucleotide, EST, GSS, Protein, Genome, and Genome Project Databases

Display Format	Description	Databases Available
Summary	Default display, hotlinked Accession number and brief description	Nucleotide, Protein, CoreNucleotide, EST, GSS, PopSet, Genome, Genome Project
Brief	Hotlinked Accession number and abbreviated description, hotlinked project number in the case of Genome Project	Nucleotide, Protein, CoreNucleotide, EST, GSS, PopSet, Genome, Genome Project
GenBank	Full report format	Nucleotide, CoreNucleotide, EST, GSS, Genome
GenPept	Full report format	Protein
GenBank(Full)	Complete GenBank record with all features, and all Sequence. This format is useful for very large GenBank records.	Nucleotide, CoreNucleotide, EST, GSS, Genome
GenPept (Full)	Complete GenPept record with all protein features, and all Sequence. This format is useful for very large GenBank records.	Protein
INSDSeq XML	XML DTD for sequence records	Nucleotide, Protein

GI list	List of GenInfo -- GI identifiers	Nucleotide, CoreNucleotide, EST, GSS Protein
ASN.1	Abstract syntax Notation One, used data storage and retrieval and to help achieve interoperability among platforms.	Nucleotide, Protein, CoreNucleotide, EST, GSS, PopSet, Genome
EST	Native display format for Expressed Sequence Tag records	EST
FASTA	The definition line and sequence characters	Nucleotide Protein
Graphics or Graph	The graphical view of the sequence accessible by selecting the hotlinked Accession numbers	Nucleotide, Protein, and Genome
GSS	Native display format for Genome Survey Sequences	GSS
TinySeq XML	Simplified XML for parsing	Nucelotide, CoreNucleotide, EST, GSS, Genome, Protein
Overview	Tabular-layout of data including Links to BLAST results, CDD, ftp site, and general information for a genome In Genomes; for Genome Project database it is a complete display of links to projects in the database, serves as a portal to links to all projects in the database about the organism-specific genome.	Genome, Genome Project
PopSet Summary	The complete set of Accession Numbers comprising the PopSet accessible by selecting the hotlinked PopSet Accession Numbers	PopSet
UI List	list of database ID's	PopSet
XML	Script-parseable format	Nucleotide Protein Genome

Selecting Documents, Displaying Them, or Accessing Their Links

Check boxes to the left of each document summary result are used to select individual documents from a set of documents retrieved. Once selected, the documents can be displayed (in various formats), sent to the Clipboard, or sent to a local disk. Documents are deselected by unclicking the check box.

Select Documents Using Check Boxes

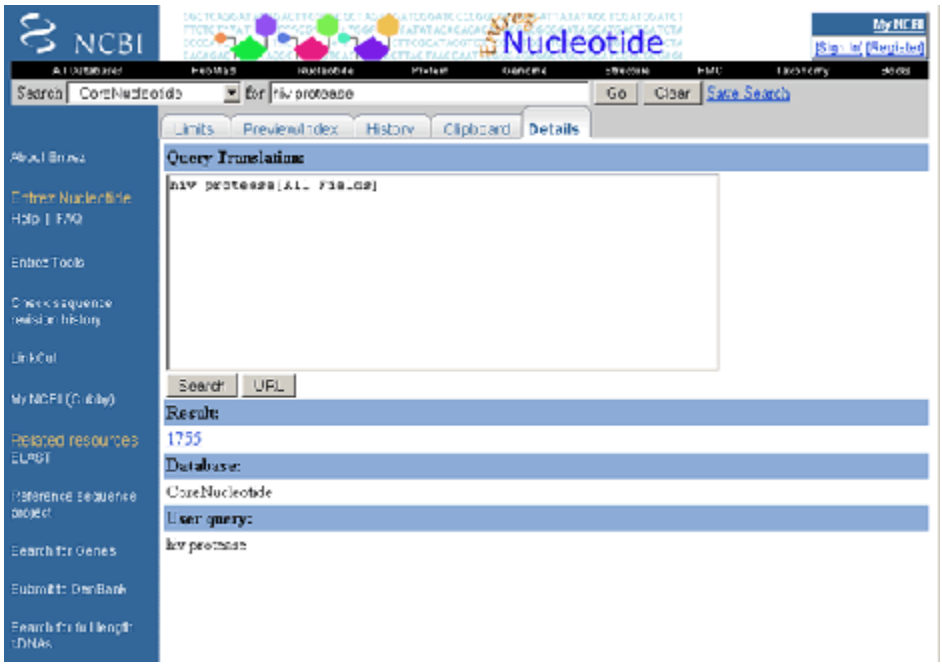
For a **useable FASTA format** that can be easily used in other applications, select the **Send to** text option. The Send to text option uses your browser to display the sequence in FASTA format. Copy and paste the sequence from the browser to other applications. Also see the section below on saving to local disk for information on saving more useable data formats from Entrez.

Details Button, Send To Clipboard, and Save

Details - click the Details tab to display your search strategy as translated using Entrez's search and syntax rules.

The Details window also contains error messages, when applicable. Note that the Details report shows the database searched, the number of documents retrieved (with hotlinks to the documents), and your search statement as written (i.e., not translated by Entrez). Within the Details window, you can modify and resubmit your search strategy. Submit the modified search query by selecting the Search button.

Details



Clipboard

The Clipboard is a temporary place to save search results. Each Entrez database has its own. Search results are not saved automatically. Each database clipboard is limited to 500 items, and items saved to the clipboard are lost after 8 hours of inactivity. Items can be displayed and saved from the Clipboard. See the Details, Send To Clipboard, and Save section of this document for help in adding records to and using records on your clipboard.

Adding to the Clipboard - select documents 1, 3, and 5 from the results set by clicking on the check box adjacent to the document number. Then click the 'Clipboard' button. Note that three items were added to the Clipboard. You are also reminded that the Clipboard is limited to 500 items and that these three items will be lost after 8 hours of inactivity during a single search session. Also, please note that the document numbers for these items (i.e., documents 1, 3, and 5) are now shown in green to indicate that they are on the Clipboard. This feature is useful because as you continue to search, if these documents are retrieved through other search strategies, their document numbers will appear in green to indicate that they are already on the Clipboard.

Retrieving documents from the Clipboard - select the Clipboard button. The items on the Clipboard are displayed in the default Summary format. Note that the documents are renumbered, but the numbers are in green to indicate that the items are on the Clipboard. Also please note that you

can display Clipboard items in all available formats, and you can link to document neighbors or related items in other databases. Items are removed from the Clipboard by selecting the items using the check box and selecting the “Clip Remove” option from the 'Send to' menu.

Saving to a local disk - select the Save button at the top (or bottom) of the results display screen next to the 'Text' button. Documents can also be saved from the Clipboard in the same manner described here. Before clicking the 'Save' button, decide two things: which documents you want and in what format. After selecting your documents by clicking on the check boxes and choosing the format using the format pull-down menu, select the 'Display' button. Once they are all displayed, click the Save button. You are prompted to name the file to which the results are saved on your local drive. If you do not select specific documents, all documents in the results set are saved. In the example below, documents 2, 3, 4, 6, and 9 will be saved to disk in the FASTA format. If these documents were not selected, all 30 documents (i.e., the entire retrieved set) would be saved to disk in the FASTA format.

Printing - use the Print function of your Web browser. As with saving to local disk, before printing, decide two things: Which documents you want to print and in what format. Because you are using the Web browser print function, you can only print documents that are displayed. Therefore, consider increasing the number of documents displayed per page so that the total number of documents you want to print are displayed on one page. Print hints: To save paper, consider using the Text or Save buttons before printing. Doing so will eliminate everything but the actual data you need (i.e., Entrez search interface, menu bars). If you use the Text button, print from your Web browser. If you use the Save button, print from another application on your machine.

LinkOut

LinkOut is a service that provides links from Entrez records to NCBI resources, such as UniGene and LocusLink, and to external resources, such as full-text journal articles, biological data, and sequence centers. These other resources provide a URL, resource name, and brief description of their Web site, which PubMed uses to create the links to their sites. User registration, a subscription fee, or some other type of fee may be required to access the full text of articles in some journals using this feature. Information for developers is available at LinkOut [http://www.ncbi.nlm.nih.gov/entrez/linkout/doc/linkoutoverview.html].

Entrez Tools For The Power User

The Advanced Entrez Tools page contains a link to Batch Entrez, which allows searchers to retrieve results from a list of existing identifiers such as GI’s or accession numbers.

Entrez Programming Utilities

The Entrez Programming Utilities (E-Utils) are a set of eight server-side programs that provide a stable interface to the Entrez query and database system. The E-Utilities use a fixed URL syntax that translates a standard set of input parameters into the values necessary for various NCBI soft-

ware components to search for and retrieve data. The E-Utilities are therefore the structured interface to the Entrez system databases. To learn more about E-Utilities, see Building Customized Data Pipelines Using the Entrez Programming Utilities (eUtils) [http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=coursework.chapter.eutils]or consider taking a course on NCBI PowerScripting [http://www.ncbi.nlm.nih.gov/Class/PowerTools/eutils/course.html]**.**